Look-Back: Implicit Visual Re-focusing in MLLM Reasoning

Shuo Yang[1*], **Yuwei Niu**[1*], **Yuyang Liu**[1†], **Yang Ye**[1], **Bin Lin**[1], **Li Yuan**[1,2†]

[1]Peking University, Shenzhen Graduate School
[2]Peng Cheng Laboratory
shuo_yang@stu.pku.edu.cn, yuanli-ece@pku.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) have achieved remarkable progress in multimodal reasoning. However, they often excessively rely on textual information during the later stages of inference, neglecting the crucial integration of visual input. Current methods typically address this by explicitly injecting visual information to guide the reasoning process. In this work, through an analysis of MLLM attention patterns, we made an intriguing observation: with appropriate guidance, MLLMs can spontaneously re-focus their attention on visual inputs during the later stages of reasoning, even without explicit visual injection. This spontaneous shift in focus suggests that MLLMs are intrinsically capable of performing visual fusion reasoning. Building on this insight, we introduce **Look-Back**, an implicit approach designed to guide MLLMs to "look back" at visual information in a self-directed manner during reasoning. Look-Back empowers the model to autonomously determine when, where, and how to re-focus on visual inputs, eliminating the need for explicit model-structure constraints or additional input. We demonstrate that Look-Back significantly enhances the model's reasoning and perception capabilities, as evidenced by extensive empirical evaluations on multiple multimodal benchmarks.

Code — https://github.com/PKU-YuanGroup/Look-Back

1 Introduction

With the development of multimodal reasoning (Gupta and Kembhavi 2023) and reinforcement learning with verifiable rewards (RLVR) (Shao et al. 2024b; Guo et al. 2025), Multimodal Large Language Models (MLLMs) (Team 2025) have made significant progress in jointly processing image and text inputs to perform complex tasks. However, recent research indicates that most approaches still predominantly rely on text during the later stages of reasoning, neglecting the visual modality (Zheng et al. 2025b; Fan et al. 2025; Yang et al. 2025b; Hu et al. 2024). Specifically, during the reasoning process, the model's attention to visual information gradually diminishes, almost reaching zero in the later stages (Tu et al. 2025; Chen et al. 2024), to the extent that visual information in the later phases exerts negligible influence on the reasoning result (Sun et al. 2025).

However, humans naturally integrate visual and cognitive processing in multimodal reasoning (Tversky, Morrison, and Betrancourt 2002), and OpenAI's o3 (OpenAI 2025) represents the gradual shift in the field from solely text-based reasoning to deep integration with visual information. Despite this progress, most existing methods still **explicitly inject visual information** (Zhang et al. 2025b; Wang et al. 2025c), such as re-inputting images or re-injecting image tokens into the model (Wu et al. 2025; Xu et al. 2025; Gupta and Kembhavi 2023). These methods essentially guide the model to re-focus its attention on visual cues. Based on this, we propose a critical research question:

Instead of explicitly re-injecting visual information, can MLLMs be enabled to self-directively and implicitly learn when and how to re-focus on visual input?

Based on the aforementioned question, we conducted a preliminary experiment to validate that the model can autonomously re-focus on the image. Specifically, we introduced a simple prompt (as shown in Figure 2) into the original CoT framework. Surprisingly, the model spontaneously enhanced its attention to the image during the later stages of reasoning, re-focusing on the visual input without any additional explicit inputs or model-structure constraints.

To better leverage the model's spontaneous attention to the image, we propose the **Look-Back** method, which guides MLLMs to naturally "look back" at visual information during reasoning, enhancing focus on visual input. Specifically, we developed a two-stage training framework: in the first stage, advanced MLLMs generate reflective data with the <back> token, followed by cold-start fine-tuning to prepare for RL. In the second stage, we apply a format reward based on the <back> token for the GRPO algorithm to strengthen the model's attention to visual input.

As shown in Figure1, Look-Back encourages MLLMs to generate reflective reasoning related to the image without explicitly injecting visual information, enhancing attention during later reasoning stages (i.e., re-focusing). Attention map analysis confirms that MLLMs attends to the correct visual locations with the <back> token. Look-Back enables MLLMs to autonomously decide when (timing of the <back> token), where (regions to focus on), and how (au-

[*]Equal contribution
[†]Corresponding author

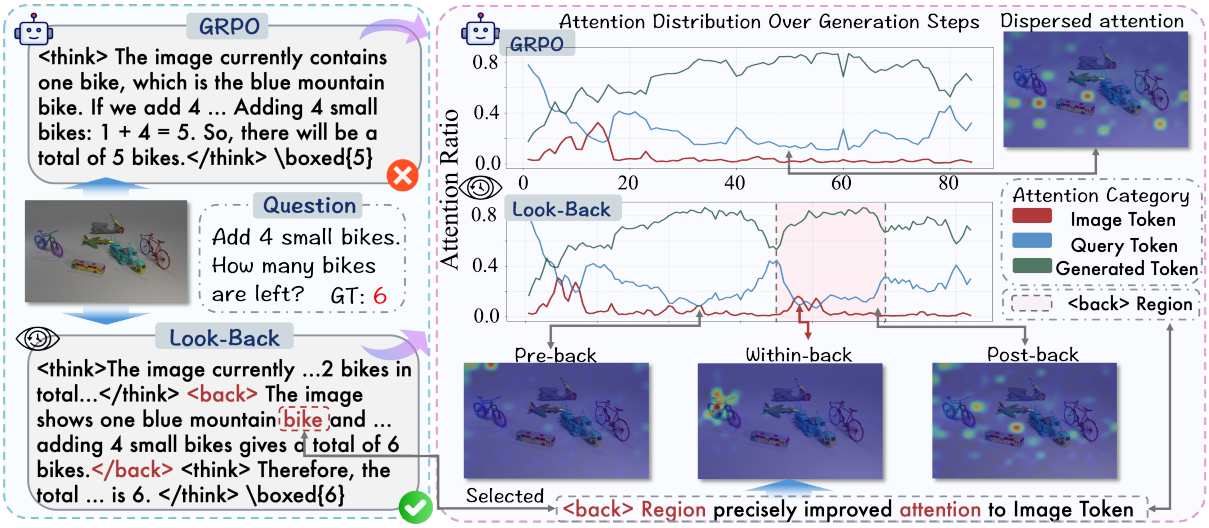

Figure 1: Overview of the Look-Back Mechanism. This figure compares the GRPO model with our Look-Back approach. The GRPO model (top left) miscounts bikes due to reduced visual attention in later stages. In contrast, the Look-Back model (bottom left) uses the <back> token to re-focus on the image, correcting the count. The attention trends show that Look-Back significantly increases attention to image tokens (red line) during the <back> phase, a behavior absent in GRPO. The attention maps below confirm that this re-focused attention is precisely targeted at the relevant objects in the image.

tonomously determining how to enhance attention) to reflect on visual input, without explicit visual inputs.

This paper proposes an implicit visual fusion reasoning paradigm, generated spontaneously by the model, rather than evaluating which paradigm is most effective. We validated this using the Qwen-2.5-VL-7B model (Team 2025) on multiple multimodal reasoning benchmarks. Results show that guiding the model to spontaneously re-focus on the image with Look-Back consistently improves performance in reasoning and perception tasks. Our key contributions are summarized as follows:

- By analyzing the trend of attention changes, we found that, without explicitly injecting visual information, the existing MLLM can autonomously attend to visual input.
- We introduced the Look-Back implicit training paradigm, which triggers the model's visual reflection behavior by modifying the format reward after cold-start fine-tuning
- Extensive evaluation on multiple multimodal benchmarks demonstrated that, Look-Back can consistently enhance performance in reasoning and perception tasks.

2 Do MLLMs Know When and How to Reflect on Visual Input?

Recent research (Hu et al. 2024; Zhang et al. 2025b; Fan et al. 2025; Zheng et al. 2025b) shows that MLLMs often over-rely on text in later inference stages, neglecting the integration of visual input. This reduced attention to visual information impacts the reliability and performance of vision-language models. Current approaches typically address this by re-inputting visual information to guide reasoning.

To investigate if MLLMs can spontaneously reactivate attention to visual inputs without external intervention, we conducted a preliminary experiment using a prompt modification that encourages the model to generate a <back> token and re-examine its response based on visual information.

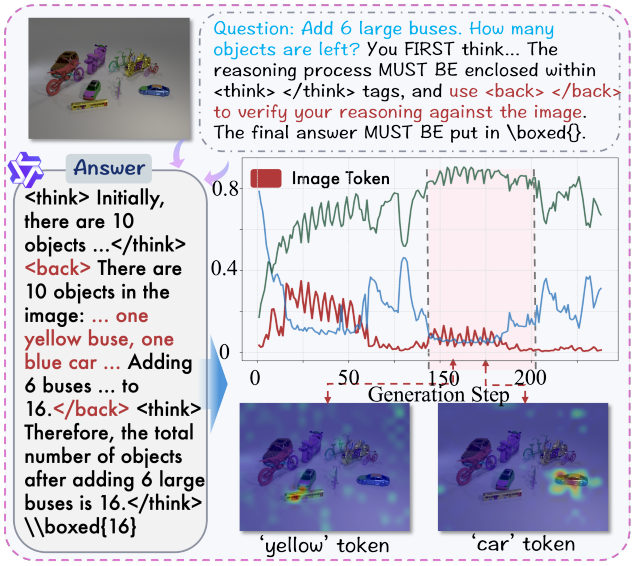

Figure 2: This figure shows how a modified prompt can encourage MLLM to spontaneously generate <back> tokens and re-examine its reasoning against visual information. This triggers the model to autonomously re-focus its attention on specific visual details, like the yellow bus and car shown in the attention maps, to verify its conclusions.

Surprisingly, as shown in Figure 2, the model demonstrates a strong capacity for spontaneous visual attention recovery. Upon generating the <back> token, the model redi-

Model	MathVerse	MathVision	MathVista	WeMath	GeoMath	Avg_M
CoT prompt	45.71	25.49	64.2	60.46	45.61	48.294
Back prompt	46.62	26.58	67.3	63.22	47.93	50.33
Trigger rate	62.01%	56.05%	87.00%	51.26%	56.10%	62.48%

Table 1: Performance of Qwen-2.5-VL-7B on Math-Benchmark with different prompts. "CoT prompt" refers to the standard Chain-of-Thought prompt. "Back prompt" encourages the model to re-focus on the image using <back> tokens. "Trigger rate" denotes the percentage of responses where the model generated the <back> token.

Model	MathVerse	MathVision	MathVista	WeMath	GeoMath	Avg_M
w/o *back*	48.67	24.93	65.29	67.26	48.3	50.89
w/ *back*	50.14	27.63	70.69	70.63	51.74	54.166
Δ *Gain*	+1.47	+2.70	+5.40	+3.37	+3.44	+3.276

Table 2: Performance on Math-Benchmark comparing models *with* and *without* the <back> mechanism, specifically for instances where the <back> token was triggered. "w/o *back*" represents the baseline performance, while "w/ *back*" shows the performance when the model engages in visual reflection. Δ *Gain* shows the percentage performance increase.

rects attention to the visual input, as evidenced by the sharp increase in the "Image Token" attention ratio in the central graph. Critically, this is not merely a general glance at the image; the model's reasoning becomes precisely grounded in the visual evidence. The attention maps below show that during the <back> sequence, the model specifically focuses on objects—such as the yellow bus when generating the yellow" token and the gold car for the car" token. This targeted refocusing occurs intrinsically, without explicit re-injection of visual information or architectural modifications.

The results in Table 1 show improvements across benchmarks, validating that MLLMs have latent capabilities for self-directed visual reflection. To verify the performance gains from the back mechanism, we analyzed the subset of questions where the "Back prompt" triggered visual reflection. As shown in Table 2, visual reflection led to even greater improvements across all benchmarks. However, the "Trigger rate"' in Table 1 reveals a limitation: modifying prompts alone is insufficient to consistently trigger reflection, with a 62.48% average trigger rate, highlighting the need for further optimization. Therefore, we propose using reinforcement learning to better incentivize this mechanism.

3 Method of Look-Back

The **Look-Back** method guides MLLMs to spontaneously re-focus on visual inputs during inference, enhancing visual fusion reasoning. It consists of two stages: supervised fine-tuning (SFT) and reinforcement learning (RL).

3.1 Cold-start Initialization

To address instability from spontaneous triggering of the <back> token and reward hacking, we created a supervised fine-tuning dataset for cold-start initialization. Based on when the <back> token is triggered, we classify the backtracking prompts into two categories:

- **Semantic-level backtracking (Semantic-back):** Triggered during reasoning, allowing the model to revisit visual details for intermediate steps and continue reasoning.
- **Solution-level backtracking (Solution-back):** Triggered after generating a preliminary solution, prompting model to reconsider visual input for a comprehensive rethink.

We designed two output formats.

Output Formats of Backtracking

Semantic-back: <think> reasoning process here </think> <**back**> verification process here <**/back**> <think> continue reasoning </think> \boxed{final answer}.

Solution-back: <think> reasoning process here </think> <**back**> verification process against the image here <**/back**> <think> based on the thinking and verification contents, rethinking here </think> \boxed{final answer}.

Data Construction. The data construction process, as illustrated in Figure 3 (A), involves three steps:

1. **Model Inference:** Perform Chain-of-Thought (CoT) inference using Qwen-2.5-VL-7B with $n = 12$ independent inferences per question.
2. **CoT Selection:** Based on the inference results, we calculate the accuracy reward and select the questions that have a higher reward variance and greater difficulty.
3. **Advanced Model Insertion:** The question, image, CoT reasoning, and ground-truth answer are input into o4-mini, which inserts backtracking tokens based on predefined rules. For correct answers, image validation tokens are added; for incorrect answers, correction tokens based on images are inserted, and the final answer is adjusted.

Supervised Fine-Tuning (SFT). Using the cold-start dataset generated with the <back> tokens, we apply SFT to guide the model to consistently trigger the backtracking behavior. Each sample is represented as $(x, q, r_{\text{back}}, a)$, where

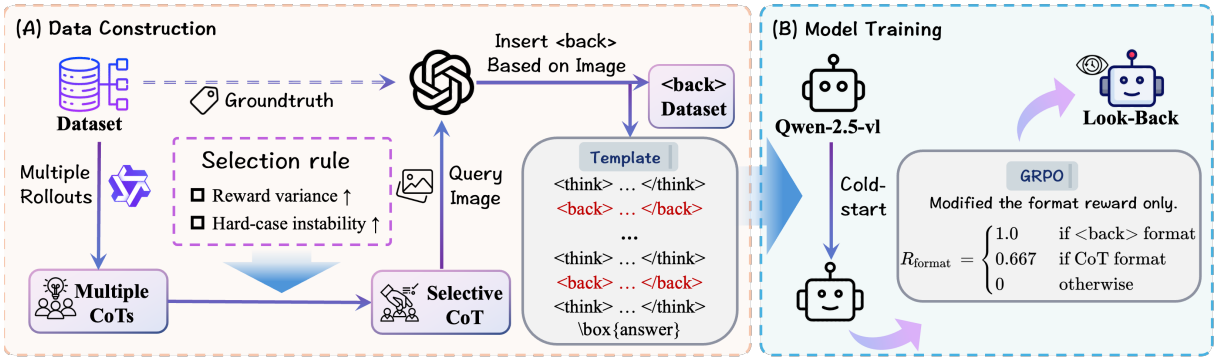

Figure 3: Pipeline of the Look-Back Method, including data construction, reflective SFT, and RL. (A) Data construction: Insertion of <back> tokens based on the image to generate a modified dataset. (B) Model training: Train Qwen-2.5-vl with cold-start and Look-Back, adjusting the reward format.

x denotes the input image, q represents the question, r_{back} is the backtracking token sequence, and a is the answer sequence. The training objective is as follows:

$$\mathcal{L}_{\text{cold-start}} = -\mathbb{E}_{(x,q,r_{\text{back}},a)\sim\mathcal{D}} \sum_{t=1}^{|y|} \log \pi_\theta \left(y_t \mid x, q, y_{<t}\right), \tag{1}$$

where $\mathcal{D}$ denotes the dataset, and $y = [r_{\text{back}}; a]$ concatenates the backtracking tokens and answer sequence.

3.2 Look-Back Reinforcement Learning

To enhance the model's ability to autonomously revisit visual inputs, we employed the Group Relative Policy Optimization (GRPO) algorithm. Compared to traditional methods, GRPO optimizes policy gradients within a sample group, enabling more diverse and efficient reasoning responses. The optimization objective is as follows:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\big[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O \mid q)\big] \\ \frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \Big\{ \min\Big[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})} A_{i,t}, \\ \text{clip}\Big(\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, 1-\varepsilon, 1+\varepsilon \Big) A_{i,t} \Big] - \beta\, \mathbb{D}_{\text{KL}}\big[\pi_\theta \,\|\, \pi_{\text{ref}}\big] \Big\}, \end{aligned} \tag{2}$$

where ϵ and β are the clipping hyperparameters and the KL divergence penalty coefficient, respectively. To guide the model in triggering the visual review behavior more stably, we modified only the format reward function. Specifically, the format reward function R_{format} is defined as:

$$R_{\text{format}} = \begin{cases} 1.0, & \text{if <back> format,} \\ 0.667, & \text{if CoT format,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The complete reward function is a combination of the format reward and accuracy reward, defined as:

$$R = \lambda \cdot R_{\text{format}} + R_{\text{accuracy}}, \tag{4}$$

where R_{accuracy} is the accuracy reward, and λ balances the format and accuracy rewards. This function motivates the model to autonomously revisit visual information, enabling active reflection during reasoning, similar to how humans naturally revisit images without explicit re-injection.

4 Look-Back Experiments Analysis

4.1 Experimental Setup

Baselines and Benchmarks. We evaluated Look-Back on eight benchmarks: five mathematical (MathVerse, MathVision, MathVista, WeMath, GeoMath) (Zhang et al. 2024; Wang et al. 2024; Lu et al. 2023; Qiao et al. 2024; Tan et al. 2025) and three perceptual (HallusionBench, TallyQA, MME) (Guan et al. 2024; Acharya, Kafle, and Kanan 2019; Fu et al. 2024). Performance was averaged per category. We compared Look-Back with three baselines: (1) Closed-Source MLLMs (e.g., GPT-4o, o3) (Hurst et al. 2024; OpenAI 2025), (2) Open-Source General MLLMs (e.g., Qwen2.5-VL-32B, InternVL3-38B) (Team 2025; Zhu et al. 2025), and (3) Open-Source Reasoning MLLMs (e.g., MM-Eureka-8B, R1-VL-7B, VL-Rethinker-7B, etc.) (Meng et al. 2025; Zhang et al. 2025a; Wang et al. 2025a; Deng et al. 2025; Wang et al. 2025b; Chen et al. 2025; Huang et al. 2025; Yang et al. 2025b; Liu et al. 2025a).

Training Datasets. For the RL phase, we used 15k mathematical problems from Geo170K, Math360K, Geometry3K, and K12 datasets (Gao et al. 2023; Shi et al. 2024; Lu et al. 2021; Meng et al. 2025). In the SFT phase, we applied the data construction from Section 3.1 to create 10k cold-start datasets for Semantic-back and Solution-back.

Implementation Details. Training was done on eight NVIDIA A800 GPUs with the Qwen2.5-VL-7B-Instruct model. SFT was performed for one epoch to prevent overfitting. For RL, we used EasyR1 (Zheng et al. 2025a) with a reward weight λ of 0.1, trained for two epochs on the 15k dataset with a batch size of 128 and a temperature of 1.0.

4.2 Main Results

Mathematical Reasoning. As shown in Table 3, our Look-Back approach outperforms the base model across all benchmarks. On five mathematical benchmarks, Semantic-back improved by 7% (from 48.5% to 55.5%), and Solution-back showed a 7.9% increase (from 48.5% to 56.4%). We also compared Look-Back with ten Open-Source Reasoning MLLMs. Despite variations in training data and duration, making direct comparison challenging, Look-Back

Model	Math-Benchmark						Perception-Benchmark				Overall
	MathVerse	MathVision	MathVista	WeMath	GeoMath	Avg_M	Hallusion	TallyQA	MME	Avg_P	Avg_{All}
Closed-Source MLLMs											
Claude-3.7	52†	41.3†	66.8†	72.6†	-	-	-	-	-	-	-
GPT-4o	50.8†	30.4†	63.8†	69†	-	-	-	-	-	-	-
GPT-o1	57†	60.3†	73.9†	98.7†	-	-	-	-	-	-	-
GPT-o3	-	-	86.8†	-	-	-	-	-	-	-	-
Gemini-2-flash	59.3†	41.3†	70.4†	71.4†	-	-	-	-	-	-	-
Open-Source General MLLMs (7B-38B)											
InternVL2.5-8B	39.5†	19.7†	64.4†	53.5†	63	48.0	61.7	53.9	-	-	-
InternVL2.5-38B	49.4†	31.8†	71.9†	67.5†	-	-	70.0	-	-	-	-
InternVL3-8B	39.8†	29.3†	71.6†	-	45.6	-	64.3	-	85.1 / 2322	-	-
InternVL3-38B	48.2†	34.2†	75.1†	-	48.2	-	72.0	75.1	87.7 / 2403	78.3	-
QwenVL2.5-7B	46.3†	25.1†	68.2†	62.1†	45.6	49.5	65.0	75.5	82.1 / 2180	74.2	61.8
QwenVL2.5-32B	48.5†	38.4†	74.7†	69.1†	54.5	57.0	71.8	79.2	88.4 / 2444	79.8	68.4
Open-Source Reasoning MLLMs (7B)											
MM-Eureka-8B	40.4†	22.2†	67.1†	58.7	50.7	47.8	65.3	76.9	84.4 / 2306	75.5	61.7
R1-VL-7B	40.0†	24.7†	63.5†	60.1	47.7	47.2	54.7	72.9	86.4 / 2376†	71.3	59.3
VL-Rethinker-7B	52.9	30.0	74.4	69.1	50.0	55.3	69.9	76.5	86.9 / 2336	77.8	66.0
OpenVLThinker-7B	45.7	26.3	71.2	66.7	55.0	53.0	70.2	80.1	86.4 / 2328	78.9	65.5
ThinkLite-VL-7B	49.3	26.2	71.7	61.9	46.5	51.1	70.7	80.3	87.6 / 2378	**79.5**	65.6
VLAA-Thinker-7B	52.7	29.2	69.7	70.2	48.8	54.1	68.2	78.2	84.8 / 2356	77.1	65.3
Vision-R1-7B	52.4†	28.0	70.6	73.9	48.6	54.7	65.5	78.1	84.4 / 2312	76.0	65.3
MM-Eureka-Qwen-7B	50.5	28.9	70.4	65.2	47.7	52.5	68.6	78.3	86.1 / 2370	77.7	65.1
R1-Onevision-7B	46.4	29.9	64.1	61.8	47.7	50.0	67.5	76.7	82.3 / 2284	75.5	62.7
NoisyRollout-7B	53.2†	27.8	72.5†	70.8	50.7	55.0	70.8	77.4	81.8 / 2038	76.7	65.8
Semantic-back-7B	50.5	27.7	71.6	71.3	56.5	**55.5**	70.7	81.2	87.1 / 2340	**79.6**	**67.6**
Solution-back-7B	51.8	30.3	72.3	70.8	56.7	**56.4**	69.8	79.2	85.9 / 2319	78.3	**67.3**

Table 3: Comparison of our Look-Back models (*Semantic-back* and *Solution-back*) with representative **Closed-Source**, **Open-Source General**, and **Open-Source Reasoning MLLMs** across the **Math-Benchmark** and **Perception-Benchmark** suites (higher is better). † scores are taken from the respective models' official reports.

Model	Math-Benchmark						Perception-Benchmark				Overall
	MathVerse	MathVision	MathVista	WeMath	GeoMath	Avg_M	Hallusion	TallyQA	MME	Avg_P	Avg_{All}
Qwen-2.5-VL-7B	46.3†	25.1†	68.2†	62.1†	45.6	49.5	65.0	75.5	82.1	74.2	61.8
+GRPO	49.3	26.8	70.9	67.6	55.2	53.9	68.6	78.3	85.5	77.5	65.7
Semantic-back-7B	**50.5**	**27.7**	**71.6**	**71.3**	**56.5**	**55.5**	**70.7**	**81.2**	**87.1**	**79.6**	**67.6**
w/o *SFT*	49.7	27.3	71.3	70.1	56.3	54.9	69.3	79.5	86.6	78.5	66.7
w/o *RL*	44.7	24.4	63.8	58.9	37.9	45.9	68.5	67.4	77.1	71.0	58.5
Solution-back-7B	**51.8**	**30.3**	**72.3**	**70.8**	**56.7**	**56.4**	**69.8**	**79.2**	85.9	**78.3**	**67.3**
w/o *SFT*	49.5	27.9	72.0	70.3	56.2	55.2	69.3	79.1	**86.0**	78.1	66.7
w/o *RL*	43.4	20.2	63.1	52.3	36.2	43.0	65.0	74.3	83.4	74.2	58.6

Table 4: Ablation of the *Look-Back*, selectively removing SFT or RL for both *Semantic-back* and *Solution-back*. Removing either phase significantly degrades performance, while both back mechanisms outperform the standard GRPO.

demonstrated competitive performance, and Solution-back, with fewer parameters, narrowed the gap with closed-source models thanks to the "look-back" mechanism.

Perceptual Reasoning. Although our training primarily used mathematical reasoning data, it is noteworthy that on perceptual benchmarks, Semantic-back improved by 6.3% (from 61.3% to 67.6%) and Solution-back showed a 6% increase (from 61.3% to 67.3%) compared to the baseline. Our approach also exhibited strong competitiveness with other Open-Source Reasoning MLLMs. These results highlight the importance of the "look-back" mechanism in enhancing the generalization of multimodal reasoning systems.

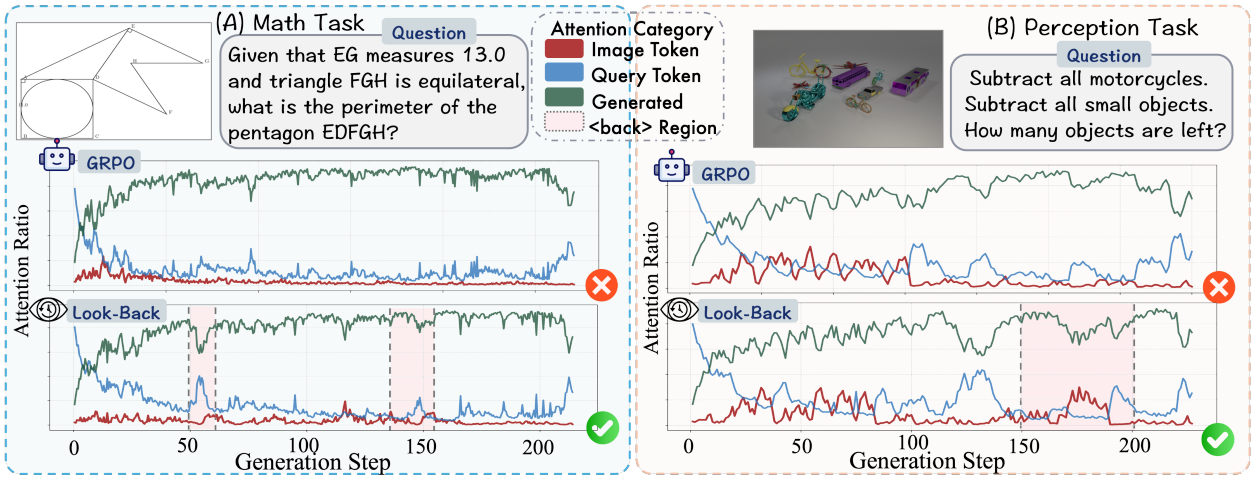

Figure 4: Look-Back Enhances Visual Grounding with Multiple Verifications. The graphs illustrate that, unlike models trained with standard GRPO, our model successfully and repeatedly re-focuses on visual input (spikes in red line) during the later reasoning stages for both Math (A) and Perception (B) tasks. This visual verification can occur multiple times within one task, demonstrating an autonomous ability to revisit and ground reasoning in visual evidence.

Model	Math-Benchmark						Perception-Benchmark				Overall
	MathVerse	MathVision	MathVista	WeMath	GeoMath	Avg_M	Hallusion	TallyQA	MME	Avg_P	Avg_All
RR-10%	49.0	26.6	69.9	68.1	56.7	54.1	68.8	80.5	86.7	78.6	66.3
RR-30%	**50.8**	**28.8**	71.5	69.9	**56.8**	**55.6**	69.5	80.3	86.1	78.6	67.1
RR-50%	50.5	27.7	**71.6**	**71.3**	56.5	55.5	**70.7**	**81.2**	**87.1**	**79.6**	**67.6**
RR-70%	50.0	27.6	70.6	69.1	55.0	54.5	68.8	79.7	86.3	78.2	66.4
RR-90%	49.9	27.2	70.7	70.0	56.6	54.9	70.4	80.6	86.9	79.3	67.1

Table 5: Effect of reflection-rate on performance. *RR-x%* denotes training with an x% reflection rate. The optimal reflection rate for tasks is between 30% and 50%, with extreme values leading to decreased performance.

4.3 Ablation Study

Effectiveness of Look-Back. We investigate the contributions of each stage within the Look-Back framework. As shown in Table 4, removing either the RL or SFT phase significantly degrades performance. Compared to standard GRPO without the look-back mechanism, both Semantic-level and Solution-level back mechanisms show performance improvements.

Ablation of Reflection Rate. The "reflection rate" refers to the SFT sampling ratio of ¡back¿ segments in the training data. Since the look-back process involves both verification and reflection-based correction, providing a single look-back dataset during the SFT cold-start phase could lead to reward hacking. We conducted an ablation study on the reflection rate of the SFT dataset using the Semantic-level back mechanism. Results in Table 5 show the optimal reflection rate for tasks is between 30% and 50%, with extreme values decreasing performance. As a result, we adopted a reflection rate of 50% in this study.

4.4 Reasoning Qualitative Analysis

Beyond the quantitative improvements across benchmarks, we conducted qualitative analyses to verify that Look-Back alters MLLM attention patterns. As shown in Figure 4, our method consistently improves attention on both mathematical and perceptual tasks. Compared to standard GRPO, Look-Back enables models to re-focus on visual input during later reasoning stages for verification.

Further analyses reveal cases from five benchmarks, showing how *Semantic-back* and *Solution-back* use Look-Back to correct errors by grounding reasoning in visual evidence. This demonstrates that Look-Back helps MLLMs autonomously determine when, where, and how to revisit visual information, moving beyond text-based reasoning. This supports our key insight: with proper guidance, MLLMs can perform visual fusion reasoning without explicit prompting.

Look-Back is effective both qualitatively and quantitatively: attention on visual tokens increases by 42.9% inside <back> vs. <think> (2.96% $\rightarrow$ 4.23%), and corrupting the visual analysis inside <back> flips many correct answers to incorrect, confirming that <back> provides essential visual grounding rather than superficial repetition.

5 Further Discussion

5.1 Failed Attempts

In our attempts to leverage the model's ability to re-focus on images, we encountered several failures, though these do not

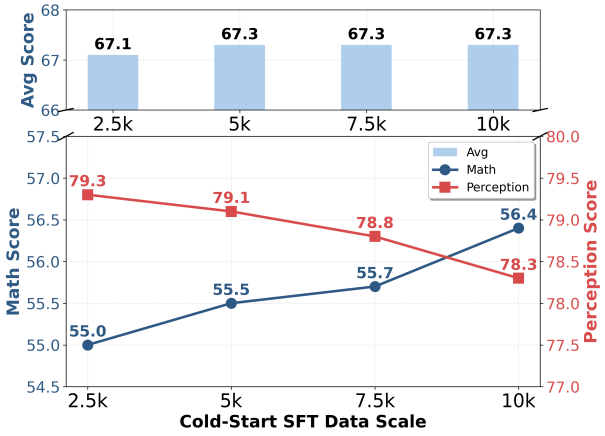

Figure 5: Impact of cold-start SFT scale (2.5k → 10k): Math scores rise steadily, Perception scores decline marginally, and the overall average remains almost flat.

imply a fundamental flaw in the approach.

Reward Hacking in Weaker Models. We initially applied Look-Back training on the Qwen-2-VL model but encountered reward hacking: the model generated empty <back></back> token sequences to gain rewards without reasoning. We suspect this is due to Qwen-2-VL lacking sufficient visual reflection capability, while Qwen-2.5-VL may have this ability due to pretraining. This aligns with prior findings (Yue et al. 2025) that reinforcement learning may fail to improve reasoning beyond the base model.

Cold-Start Data Requirements. We initially generated CoTs using GPT-4o and inserted <back> tokens but observed performance drop after cold-starting. We then used model-generated data with refined <back> insertion, improving performance. We hypothesize that fine-tuning on homologous outputs reduces distributional deviation, aligning better with the cold-start objective.

5.2 Impact of Cold Start

Scaling Cold-Start Data. To assess the effect of cold-start data scale, we experimented with 2.5k, 5k, 7.5k, and 10k mathematical samples. As shown in Figure 5, increasing cold-start data improved performance on mathematical tasks, but slightly reduced performance on perceptual tasks. However, the overall performance remained relatively unchanged. We hypothesize that cold starting with purely mathematical data may limit further generalization on perceptual tasks. Incorporating more diverse SFT and RL data could further enhance overall robustness.

Performance Differences Between Semantic-Back and Solution-Back. As shown in Table 4, both <back> methods improve performance. Semantic-back excels in perceptual tasks, while Solution-back performs better on mathematical tasks. We speculate that early backtracking facilitates timely confirmation of visual cues, benefiting perceptual tasks. In contrast, deferring backtracking until after CoT reasoning enables more comprehensive verification with minimal disruption to the reasoning chain, favoring mathematical tasks.

6 Related Work

Multimodal complex reasoning has advanced in four stages: early explicit module exploration, supervised fine-tuning and test-time scaling, reinforcement learning advancements, and the evolution of visual thinking.

Early Development of Multimodal Reasoning. Early MLLM development focused on explicit prompts and multi-module cooperation. Techniques like Visual-CoT (Shao et al. 2024a) used reasoning chains and visual sampling for reasoning. Visual-SketchPad (Hu et al. 2024) introduced a three-stage workflow with visual sketching for interpretability. Multimodal-CoT (Zhang et al. 2023) proposed a two-stage framework that separates reasoning chain generation from answer inference.

Supervised Fine-Tuning and Test-Time Scaling (Liu et al. 2025b; OpenAI 2025; Yang et al. 2025a, 2024). With models like OpenAI o1, supervised fine-tuning (SFT) using large-scale synthetic CoT data became mainstream, shifting from module-based to data-driven approaches. For instance, Virgo (Du et al. 2025) adjusts reasoning depth using varying CoTs. LLaVA-CoT (Xu et al. 2024) employs a structured reasoning template for multi-step processes. TACO (Ma et al. 2024) uses programming for tool invocation through SFT. Test-Time Scaling (TTS) (Muennighoff et al. 2025) further enhances reasoning without updates.

Reinforcement Learning Breakthroughs. The success of DeepSeek-R1 (Guo et al. 2025) marked the entry of complex reasoning into reinforcement learning fine-tuning (RFT). In the multimodal domain, DIP-R1 (Park et al. 2025) explored fine-grained image processing, while Perception-R1 (Yu et al. 2025) encoded image patches, integrating testing-time augmentation with RFT training. MM-Eureka (Meng et al. 2025) advanced visual reasoning through rule-based rewards.

Evolution of Visual Thinking. Recent research trends indicate that multimodal complex reasoning not only requires "thinking in language" but also necessitates "thinking in images". DyFo (Li et al. 2025b) simulates dynamic human visual search, and DeepEyes (Zheng et al. 2025b) enables visual-textual reasoning through end-to-end reinforcement learning. MVoT (Li et al. 2025a) alternates between text and images to complement linguistic reasoning.

Unlike explicit visual re-injection approaches (e.g., DeepEyes, MVoT) or text-only reflection (e.g., VL-Rethinker), Look-Back leverages the observation that MLLMs can spontaneously re-focus on visual inputs.

7 Conclusion

In this work, we observed that Multimodal Large Language Models (MLLMs) can autonomously re-focus their attention on visual inputs during reasoning, without explicit visual information injection. Building on this insight, we introduced the Look-Back approach, which empowers MLLMs to self-direct visual reflection through a two-stage training process combining supervised fine-tuning and reinforcement learning. Our experiments show that Look-Back significantly enhances multimodal reasoning capabilities, achieving competitive results across multiple benchmarks.

Acknowledgements

This work was supported by the China Postdoctoral Science Foundation under Grant Number BX20240013 and 2024M760113, the Natural Science Foundation of China (No. 62332002, 62425101), and Shenzhen Science and Technology Program (KQTD20240729102051063).

References

Acharya, M.; Kafle, K.; and Kanan, C. 2019. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8076–8084.

Chen, H.; Tu, H.; Wang, F.; Liu, H.; Tang, X.; Du, X.; Zhou, Y.; and Xie, C. 2025. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*.

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, 19–35. Springer.

Deng, Y.; Bansal, H.; Yin, F.; Peng, N.; Wang, W.; and Chang, K.-W. 2025. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*.

Du, Y.; Liu, Z.; Li, Y.; Zhao, W. X.; Huo, Y.; Wang, B.; Chen, W.; Liu, Z.; Wang, Z.; and Wen, J.-R. 2025. Virgo: A Preliminary Exploration on Reproducing o1-like MLLM. *arXiv preprint arXiv:2501.01904*.

Fan, Y.; He, X.; Yang, D.; Zheng, K.; Kuo, C.-C.; Zheng, Y.; Narayanaraju, S. J.; Guan, X.; and Wang, X. E. 2025. GRIT: Teaching MLLMs to Think with Images. *arXiv preprint arXiv:2505.15879*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; Wu, Y.; and Ji, R. 2024. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. arXiv:2306.13394.

Gao, J.; Pi, R.; Zhang, J.; Ye, J.; Zhong, W.; Wang, Y.; Hong, L.; Han, J.; Xu, H.; Li, Z.; et al. 2023. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14375–14385.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Gupta, T.; and Kembhavi, A. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14953–14962.

Hu, Y.; Shi, W.; Fu, X.; Roth, D.; Ostendorf, M.; Zettlemoyer, L.; Smith, N. A.; and Krishna, R. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*.

Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Xu, Z.; Hu, Y.; and Lin, S. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Li, C.; Wu, W.; Zhang, H.; Xia, Y.; Mao, S.; Dong, L.; Vulić, I.; and Wei, F. 2025a. Imagine while Reasoning in Space: Multimodal Visualization-of-Thought. *arXiv preprint arXiv:2501.07542*.

Li, G.; Xu, J.; Zhao, Y.; and Peng, Y. 2025b. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9098–9108.

Liu, X.; Ni, J.; Wu, Z.; Du, C.; Dou, L.; Wang, H.; Pang, T.; and Shieh, M. Q. 2025a. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*.

Liu, Y.; Hong, Q.; Huang, L.; Gomez-Villa, A.; Goswami, D.; Liu, X.; van de Weijer, J.; and Tian, Y. 2025b. Continual Learning for VLMs: A Survey and Taxonomy Beyond Forgetting. *arXiv preprint arXiv:2508.04227*.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Lu, P.; Gong, R.; Jiang, S.; Qiu, L.; Huang, S.; Liang, X.; and Zhu, S.-C. 2021. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*.

Ma, Z.; Zhang, J.; Liu, Z.; Zhang, J.; Tan, J.; Shu, M.; Niebles, J. C.; Heinecke, S.; Wang, H.; Xiong, C.; et al. 2024. TACO: Learning Multi-modal Action Models with Synthetic Chains-of-Thought-and-Action. *arXiv preprint arXiv:2412.05479*.

Meng, F.; Du, L.; Liu, Z.; Zhou, Z.; Lu, Q.; Fu, D.; Han, T.; Shi, B.; Wang, W.; He, J.; Zhang, K.; Luo, P.; Qiao, Y.; Zhang, Q.; and Shao, W. 2025. MM-Eureka: Exploring the Frontiers of Multimodal Reasoning with Rule-based Reinforcement Learning. arXiv:2503.07365.

Muennighoff, N.; Yang, Z.; Shi, W.; Li, X. L.; Fei-Fei, L.; Hajishirzi, H.; Zettlemoyer, L.; Liang, P.; Candès, E.; and Hashimoto, T. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

OpenAI. 2025. OpenAI o3 and o4-mini System Card. https://openai.com/index/o3-o4-mini-system-card/. Accessed: 2025-04-18.

Park, S.; Kim, H.; Kim, J.; Kim, S.; and Ro, Y. M. 2025. DIP-R1: Deep Inspection and Perception with RL Looking Through and Understanding Complex Scenes. *arXiv preprint arXiv:2505.23179*.

Qiao, R.; Tan, Q.; Dong, G.; Wu, M.; Sun, C.; Song, X.; GongQue, Z.; Lei, S.; Wei, Z.; Zhang, M.; et al. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Shao, H.; Qian, S.; Xiao, H.; Song, G.; Zong, Z.; Wang, L.; Liu, Y.; and Li, H. 2024a. Visual cot: Advancing multimodal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37: 8612–8642.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shi, W.; Hu, Z.; Bin, Y.; Liu, J.; Yang, Y.; Ng, S.-K.; Bing, L.; and Lee, R. K.-W. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*.

Sun, H.-L.; Sun, Z.; Peng, H.; and Ye, H.-J. 2025. Mitigating visual forgetting via take-along visual conditioning for multi-modal long cot reasoning. *arXiv preprint arXiv:2503.13360*.

Tan, H.; Ji, Y.; Hao, X.; Lin, M.; Wang, P.; Wang, Z.; and Zhang, S. 2025. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*.

Team, Q. 2025. Qwen2.5-VL.

Tu, C.; Ye, P.; Zhou, D.; Bai, L.; Yu, G.; Chen, T.; and Ouyang, W. 2025. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *arXiv preprint arXiv:2503.08342*.

Tversky, B.; Morrison, J. B.; and Betrancourt, M. 2002. Animation: can it facilitate? *International journal of human-computer studies*, 57(4): 247–262.

Wang, H.; Qu, C.; Huang, Z.; Chu, W.; Lin, F.; and Chen, W. 2025a. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*.

Wang, K.; Pan, J.; Shi, W.; Lu, Z.; Ren, H.; Zhou, A.; Zhan, M.; and Li, H. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37: 95095–95169.

Wang, X.; Yang, Z.; Feng, C.; Lu, H.; Li, L.; Lin, C.-C.; Lin, K.; Huang, F.; and Wang, L. 2025b. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*.

Wang, Y.; Wang, S.; Cheng, Q.; Fei, Z.; Ding, L.; Guo, Q.; Tao, D.; and Qiu, X. 2025c. Visuothink: Empowering lvlm reasoning with multimodal tree search. *arXiv preprint arXiv:2504.09130*.

Wu, J.; Guan, J.; Feng, K.; Liu, Q.; Wu, S.; Wang, L.; Wu, W.; and Tan, T. 2025. Reinforcing Spatial Reasoning in Vision-Language Models with Interwoven Thinking and Visual Drawing. *arXiv preprint arXiv:2506.09965*.

Xu, G.; Jin, P.; Hao, L.; Song, Y.; Sun, L.; and Yuan, L. 2024. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Xu, Y.; Li, C.; Zhou, H.; Wan, X.; Zhang, C.; Korhonen, A.; and Vulić, I. 2025. Visual Planning: Let's Think Only with Images. *arXiv preprint arXiv:2505.11409*.

Yang, S.; Ning, K.-P.; Liu, Y.-Y.; Yao, J.-Y.; Tian, Y.-H.; Song, Y.-B.; and Yuan, L. 2024. Is Parameter Collision Hindering Continual Learning in LLMs? *arXiv preprint arXiv:2410.10179*.

Yang, S.; Zhang, Q.; Liu, Y.; Huang, Y.; Jia, X.; Ning, K.; Yao, J.; Wang, J.; Dai, H.; Song, Y.; et al. 2025a. AsFT: Anchoring Safety During LLM Fine-Tuning Within Narrow Safety Basin. *arXiv preprint arXiv:2506.08473*.

Yang, Y.; He, X.; Pan, H.; Jiang, X.; Deng, Y.; Yang, X.; Lu, H.; Yin, D.; Rao, F.; Zhu, M.; et al. 2025b. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Yu, E.; Lin, K.; Zhao, L.; Yin, J.; Wei, Y.; Peng, Y.; Wei, H.; Sun, J.; Han, C.; Ge, Z.; et al. 2025. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*.

Yue, Y.; Chen, Z.; Lu, R.; Zhao, A.; Wang, Z.; Song, S.; and Huang, G. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Zhang, J.; Huang, J.; Yao, H.; Liu, S.; Zhang, X.; Lu, S.; and Tao, D. 2025a. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*.

Zhang, R.; Jiang, D.; Zhang, Y.; Lin, H.; Guo, Z.; Qiu, P.; Zhou, A.; Lu, P.; Chang, K.-W.; Qiao, Y.; et al. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, 169–186. Springer.

Zhang, X.; Gao, Z.; Zhang, B.; Li, P.; Zhang, X.; Liu, Y.; Yuan, T.; Wu, Y.; Jia, Y.; Zhu, S.-C.; et al. 2025b. Chain-of-Focus: Adaptive Visual Search and Zooming for Multimodal Reasoning via RL. *arXiv preprint arXiv:2505.15436*.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Zheng, Y.; Lu, J.; Wang, S.; Feng, Z.; Kuang, D.; and Xiong, Y. 2025a. EasyR1: An Efficient, Scalable, Multi-Modality RL Training Framework. https://github.com/hiyouga/EasyR1.

Zheng, Z.; Yang, M.; Hong, J.; Zhao, C.; Xu, G.; Yang, L.; Shen, C.; and Yu, X. 2025b. DeepEyes: Incentivizing" Thinking with Images" via Reinforcement Learning. *arXiv preprint arXiv:2505.14362*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.